Megan Howard

1. Introduction

The primary objective of this study was to explore a model selection procedure that balances accuracy and fairness in machine learning models. Specifically, the comparison between standard machine learning models and fairness-aware models was conducted using criteria focused on accuracy, fairness, and a combined measure of both. The task was approached through a series of methodical experiments, utilising logistic regression models with varying hyperparameters (C and solver selection) throughout.

1.1. Evaluating the Original Dataset

1.1.1 The highest accuracy model

The analysis revealed that for standard models, the optimal hyperparameter for accuracy (C value between 10^{-3} and 10^{-2}) differed from that for fairness. This observation underpins the inherent trade-off between model complexity and generalization capability. Notably, the solver type's impact on accuracy diminished beyond this optimal C value threshold, indicating a convergence in model performance.

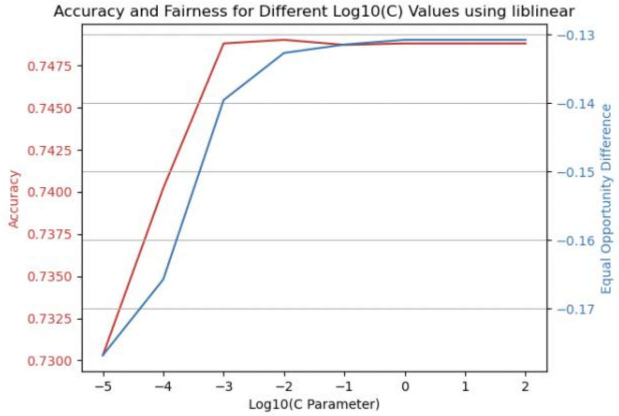

Figure 1: The effect of modifying the log(C) hyperparameter on Accuracy and EOP metrics.

1.1.2 The model with the best fairness metric

When fairness (measured by Equality of Opportunity) was prioritized, a higher C value (around 10) emerged as optimal. This suggests that fairness considerations require a different calibration of model complexity, as compared to accuracy-focused models. Additionally, an interesting observation was the initial variance in fairness performance across different solvers at lower C values, which later stabilized.

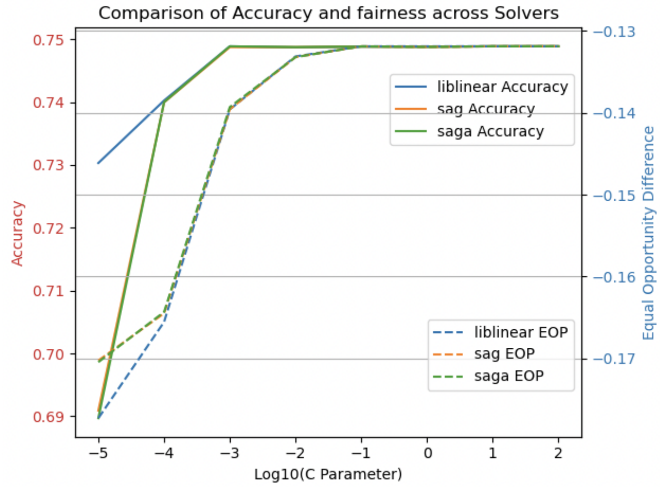

Figure 2: The effect of modifying log(C) & solver hyperparameters on Accuracy and EOP metrics for standard model method.

2. Evaluating the Fairness-Aware Dataset

For this section I used the reweighing approach of Kamiran and Calders 2012 [1].

2.1. The highest accuracy model

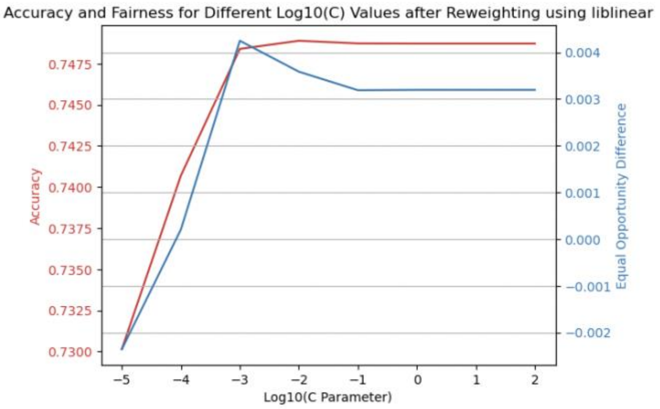

Figure 3: The effect of modifying the log(C) hyperparameter on Accuracy and EOP metrics for fairness-aware method.

Implementing the reweighing approach notably did not significantly impact the average accuracy, indicating that fairness interventions can be integrated without substantial accuracy trade-offs. This consistency in accuracy, regardless of solver type at higher C values, aligns with findings from the standard model analysis.

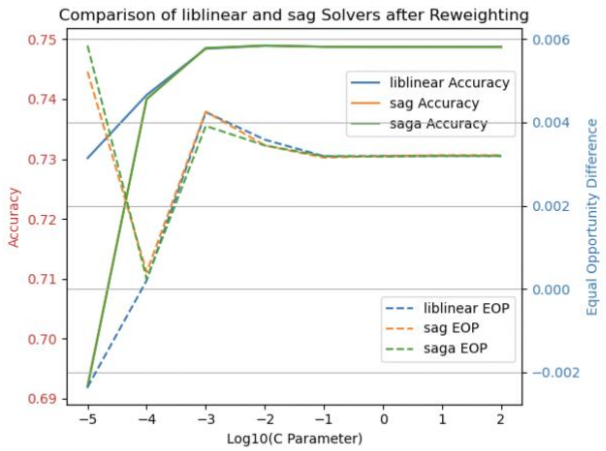

Figure 4: The effect of modifying log(C) & solver hyperparameters on Accuracy and EOP metrics for fairness- aware method.

2.2. The model with the best fairness metric

The most striking effect was observed in the Equality of Opportunity metric, which displayed a significantly reduced range in the reweighted models. The range narrowed to 0.08 (a substantial decrease from the standard model's range of 4), with the most optimal value being essentially 0. This indicates a pronounced reduction in bias, affirming the effectiveness of the fairness-aware approach in enhancing model fairness.

The results highlight a nuanced interplay between model accuracy and fairness (Table 1). Standard models, while achieving higher accuracy (0.75309), displayed larger disparities in fairness metrics (EOP around -0.61785). In contrast, fairness-aware models demonstrated a commendable balance, with a slight decrease in accuracy (0.72029) but substantial improvements in fairness (EOP -0.03176). The Pareto efficiency criterion, applied in the next section further validated these observations.

3. Model Selection Criterion to account for both Accuracy and Fairness

In selecting the most suitable model for the final task, the Pareto Efficiency approach [2,3] emerged as a superior criterion over my initially considered composite score method (Table 1). The rationale for this shift was in the nature of the Pareto Efficiency principles objectives, where multidimensional optimisation – accuracy and fairness, in this case – were both proritised, rather than one metric alone.

The initial composite scoring method I tried, had a weighted average with a dominant 80% focus on fairness and 20% on accuracy. This was decided based on the presumption that fairness should be given greater importance in the sensitive context of employment prediction based on disability status. While this method provided a singular scalar value representing a combination of accuracy and fairness, it inadvertently led to the selection of models that were highly fair but not necessarily the most accurate. Such a heavy weighting on fairness, although well-intentioned, potentially compromised the overall model effectiveness, particularly in scenarios where accuracy is also a critical factor.

Contrastingly, the Pareto Efficiency approach offered a more nuanced and equitable balance between the two desired attributes. By identifying models where no other model is both more accurate and fairer. This approach does not force an arbitrary weighting to accuracy or fairness, but rather, it highlights models that offer the best possible combinations between these two metrics. This aligns well with the structure of our output distributions seen up until this point, and also balances the ethical considerations of our task.

Upon applying the Pareto Efficiency criterion, the final selected models for both standard and fairness-aware approaches demonstrated a more balanced performance. For instance, the best standard model showed an accuracy of 0.75239 and an EOP of -0.65693, while the best fairness-aware model exhibited an accuracy of 0.72029 with an EOP of -0.03176. These results, when compared to those derived from the initial scoring method, underscore the effectiveness of the Pareto model selection strategy in accounting for a more harmonious balance between accuracy and fairness.

References

[1] Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst* **33**, 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8

[2] Corporate Finance Institute. "Pareto Efficiency, https://corporatefinanceinstitute.com/resources/economics/pareto-efficiency/ "

[3] Economics Discussion "Top Three Marginal Conditions for Pareto Optimality (With Diagram)

Model Type	**Model Parameters**	**Accuracy**	**Equal Opportunity Difference**
Best Accuracy Model - Standard	C=100.0, Solver=saga	0.75309	-0.61785
Best Fairness-aware Model - Standard	C=1.0, Solver=liblinear	0.75309	-0.61789
Best Accuracy Model - Reweighted	C=0.01, Solver=sag	0.71903	-0.00144
Best Fairness-aware Model - Reweighted	C=0.0001, Solver=saga	0.72029	-0.03176
Best Standard Model - Pareto Score	C: 0.001, Solver: liblinear	0.75239	-0.65693
Best Reweighted Model - Pareto Score	C: 0.0001, Solver: saga	0.72029	-0.03176

Table 1: Comparison of Model Performance Across Different Selection Criteria.